Covariate-Related Structure Extraction from Paired Data

Linfei Zhou[1], Elisabeth Georgii[2], Claudia Plant[3] and Christian Böhm[1]

[1] Ludwig-Maximilians-Universität München, Munich, Germany
{zhou, boehm}@dbs.ifi.lmu.de
[2] Helmholtz Zentrum München, Neuherberg, Germany
elisabeth.georgii@helmholtz-muenchen.de
[3] University of Vienna, Vienna, Austria
claudia.plant@univie.ac.at

Abstract. In the biological domain, it is more and more common to apply several high-throughput technologies to the same set of samples. We propose a Covariate-Related Structure Extraction approach (CRSE) that explores relationships between different types of high-dimensional molecular data (views) in the context of sample covariate information from the experimental design, for example class membership. Real-world data analysis with an initial pipeline implementation of CRSE shows that the proposed approach successfully captures cross-view structures underlying multiple biologically relevant classification schemes, allowing to predict class labels to unseen examples from either view or across views.

1 Introduction

With the development of modern omics technologies, massive numbers of variables are measured at the same time. For instance, sequencing technologies allow to quantify expression levels for tens of thousands of transcripts. Furthermore, multiple measurement types are frequently co-applied, providing different views on the same biological samples. Multi-view data occur also in other domains, for instance, textual descriptions combined with images of objects. In addition, biological samples often have covariate information attached, which stems from the experimental design. This information can be in form of categorical class labels such as the disease group of a patient or in form of numerical variables such as body weight. The approach proposed in this paper takes covariate information into account while analyzing relationships between different data views. The goal is to find such relationships that capture covariate-related structure in the data, for example class separation.

To integrate data from multiple views, a lot of approaches have been proposed, also known as data fusion or multi-block analysis methods (Westerhuis et al. (1998); Smilde et al. (2003); Lanckriet et al. (2004); Tenenhaus and Vinzi (2005); Jiang et al. (2012); Acar et al. (2014)). These methods have

been used in various areas (Jamali and Ester (2010); Acar et al. (2012); Lee et al. (2012)). Multi-block analysis handles multiple blocks of data collected on the same set of samples (Westerhuis et al. (1998); Smilde et al. (2003)). The main objective of multi-block analysis approaches is to find latent variables that explain each block while optimizing the correlation between blocks. The multi-block analysis methods can be classified into three groups: generalized Principal Component Analysis (PCA), Partial Least Squares (PLS) regression and Canonical Correlation Analysis (CCA) methods (Westerhuis et al. (1998); Zhou et al. (2015)). Consensus PCA (CPCA) (Svante et al. (1987)), hierarchical PCA (HPCA)(Wold et al. (1996)), multi-group multi-block PCA (mgmbPCA) (Eslami et al. (2014)) and multiple factor analysis (MFA) (Abdi et al. (2013)) are parts of multi-block family of PCA extensions introducing the concept of using multiple blocks in PCA, which identifies orthogonal directions of largest variance. PLS aims to explain an output dataset based on an input dataset (Geladi and Kowalski (1986)). A PLS approach to multi-block analysis (PLS-MBA) (Tenenhaus and Vinzi (2005)) has been proposed by Tenenhaus and Vinzi. Choi et al. also propose a multi-block PLS (MBPLS) (Choi and Lee (2005)) method as a fault detection and identification approach. CCA (Hotelling (1936); Sweeney et al. (2013); Klami et al. (2013)) is a well-known and widely used method for finding a reciprocal relationship and capturing the common variation between two datasets (Hardoon et al. (2004); Witten et al. (2009)). To handle more than two datasets, many variations of CCA methods have been proposed, such as generalized CCA (gCCA) (Horst (1961); Vía et al. (2007)) and tensor CCA (TCCA) (Luo et al. (2015)).

Related to our goal, there exist previous multi-view approaches that take covariate information into account, such as MultiwayCCA (Huopaniemi et al. (2010)) and Supervised CCA (SCCA) (Witten and Tibshirani (2009); Guo et al. (2013)). MultiwayCCA extends multiway ANOVA to the multi-view case by introducing a Bayesian model, which uses shared variables to describe common variation between both data views. SCCA searches for correlations between the data views that are associated with covariate information. As an extension of mgPCA (Krzanowski (1984)), mgmbPCA seeks common vectors of loadings across classes for each view of variables basing on an iterative algorithm. MultiwayCCA has been designed to deal with datasets where the number of variables is much larger than the number of samples (small-n-large-p problem). However, the dimension reduction step used by MultiwayCCA is very time-consuming for high-dimensional omics datasets. SCCA has a sparsity criterion that allows to deal with small-n-large-p situation, but in practice it is still slow on the high-dimensional data.

To extend classical CCA for the small-n-large-p problem, there are two principal directions, dimension reduction and penalty approaches, and the second one includes ridge regularization and sparse regularization (Vinod (1976); Saunders et al. (1998)). Regularization methods introduce additional parameters, which

leads a new problem, how to set values for these parameters. González uses a standard cross-validation procedure to choose optimal parameters (González et al. (2008)). Huopaniemi et al. employ clustering-type factor analysis as a dimension reduction technique to project the high-dimensional data into latent variables spaces (Huopaniemi et al. (2010)). Like Huopaniemi et al., we apply dimension reduction to tackle the small-n-large-p problem, but as a difference we use the covariate information already in that step. In that way we hope to solve the small size problem efficiently and at the same time effectively capture the covariate-related structure of the data.

The paper is organized as follows. The next section shows a specific pipeline implementation as a simple example of our approach, Covariate-Related Structure Extraction (CRSE). Then the effectiveness of the pipeline is demonstrated by good classification accuracy in several biological applications, which is outperforming other approaches. The final section discusses conclusions and future work.

2 Covariate information related structure extraction from multi-view data

In this section we describe a simple pipeline implementation to illustrate the idea of CRSE. The pipeline consists of two parts: covariate-dependent dimension reduction and canonical correlation analysis. The purpose of the covariate-dependent dimension reduction is on the one hand to solve the small sample size problem for the subsequent multi-view analysis and on the other hand to optimally preserve the class separation or variation of the covariate. Here we use PLS to do this. PLS projects original variables into a latent variable space that maximally explains the covariate information, hence it can reduce high-dimensional data while taking account of covariate information. PLS is applied on each data view separately. On the dimension-reduced data, CCA is applied to extract correlated structure between the data views. Figure 1 illustrates the work-flow of the proposed pipeline. The trained model can be used to project data into a covariate-dependent shared representation of both data views. We evaluate the model by classification of left-out samples in the projected space.

2.1 Partial Least Squares

PLS is a supervised method that simultaneously performs dimension reduction of the input data and regression of the output data (Boulesteix and Strimmer (2007)). In the CRSE pipeline, PLS is applied to each data view separately, using the view as input data and the covariate information as output data. The general underlying PLS model is as follows (Geladi and Kowalski (1986)):

$$X = TP^\top + E$$
$$L = UQ^\top + F^*$$

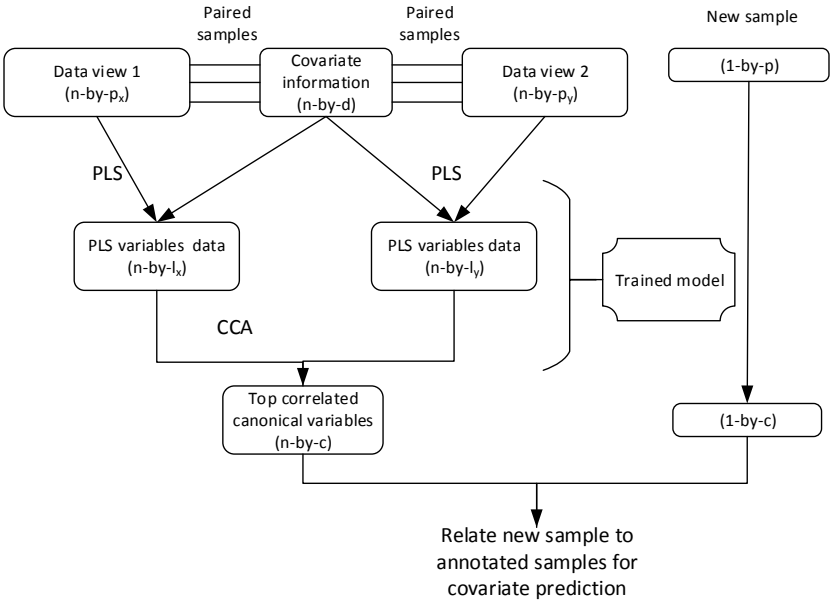

Fig. 1: The diagram of the proposed CRSE pipeline. Each data view is integrated with covariate information by PLS, yielding a low-dimensional representation of the data. The PLS variables are then further processed by CCA to find latent variables capturing shared variation between the views. The standard CCA used here handles two-view data but can be extended to multiple views.

where X is the $n \times p_x$ matrix representing the data view with original variables and L is the $n \times d$ matrix of covariate data. T and U are $n \times l$ matrices with latent variable representations of the samples, where l is fixed to a small number ($l < n < p_x$). E and F^* are error terms. The overall relation is $L = TBQ^\top + F$, where $\|F\|$ is to be minimized and B relates T and U through regression. By using PLS, we obtain covariate-aware low-dimensional representations of the two views.

2.2 Canonical Correlation Analysis

After applying PLS, the two data views are represented by low-dimensional datasets, which are used in the following step of the pipeline: an $n \times l_x$ matrix T_x and an $n \times l_y$ matrix T_y, where l_x and l_y are the number of latent PLS variables for X and Y, respectively. In the next step of the CRSE pipeline, the relationship between the two low-dimensional data views is analyzed, looking for common variation. Since l_x and l_y do not exceed the sample size (Haenlein and Kaplan (2004)), we can use for that purpose standard CCA without any regularization.

The objective of CCA is to find projections with maximal correlation between the two data views. To obtain the first CCA component, the following objective is solved (Hardoon et al. (2004)):

$$\arg \max_{a^\top T_x^\top T_x a = b^\top T_y^\top T_y b = 1} cor(T_x a, T_y b)$$

where a and b are weight vectors of length l_x and l_y. The subsequence components have the additional constraint that they are uncorrelated to earlier components. This results in two loading matrices A and B of size $l_x \times h$ and $l_y \times h$, where h is the number of paired canonical variables. Thus we get the following canonical variables:

$$C_x = T_x A$$
$$C_y = T_y B$$

Since we are interested in shared variation of the two data views, we focus on the top canonical variables to represent the samples. In practice we choose the canonical variables with a correlation greater than a threshold to focus on what is most common between datasets.

2.3 Data covariate prediction

CRSE integrates covariate information with shared components of the data views via two projection steps. Any new sample where we have either one of

the data views available can be scaled and projected by the trained PLS plus CCA model. In the projected space, we can apply a classification or regression method to predict the covariate of the new sample, allowing to assess whether the projected representation captures relevant information from the samples. Remarkably, when predicting the covariate label of a new sample in a data view, the common space allows to not only use the labeled samples of this data view but also all the labels of the other data view, including in particular non-paired labeled samples.

3 Experiments and Results

To demonstrate the effectiveness and efficiency of the CRSE pipeline, we make experiments on real-world datasets from the model plant *Arabidopsis thaliana*. Each dataset consists of two data views, a metabolomic data view and a gene expression data view. As covariate information, the biological samples are annotated by two different classification schemes, genotype and environmental condition. Genes modified in the genotypes, allowing a trivial genotype separation, were excluded from the expression data. Table 1 summarizes the key properties of both datasets.

The datasets are preprocessed using the R packages limma (Smyth (2004); Ritchie et al. (2015)), FTICRMS (Barkauskas (2012)) and nlme (Pinheiro and Bates (2000)). The CRSE pipeline is implemented based on built-in functions in R. All the experiments are executed on a regular workstation PC with 3.4 GHz dual core CPU equipped with 32 GB RAM.

Table 1: Paired datasets

	Dataset 1	Dataset 2
Nr. of paired observations	57	23
Nr. of variables in metabolomic view	1454	203
Nr. of variables in expression view	24603	24603
Nr. of classes in genotype covariate	3	2
Nr. of classes in condition covariate	6	2

The canonical variables are the projections of the original data in a new space that represents the maximum correlation structure between views and preserves covariate variation. Assuming that canonical variables keep the principal information and the basic structure of the original data, the classification result of the sample objects in the new space should have a similar accuracy to that of original data. Since irrelevant information and noise might be cleaned out, the accuracy in the new space could be even better. So we use a classification-based method to evaluate the effectiveness of the canonical variables and check whether

the detected cross-view relationships are meaningful.

For the analysis we consider paired samples between expression data and metabolome data. After applying scaling to make each variable in the training dataset to have a mean of zero and a standard deviation of one, we reuse the scaling parameters on test samples. Since we have categorical covariate information, we employ dummy coding, which uses only ones and zeros to convert all label information (Wendorf (2004); Boulesteix and Strimmer (2007)). Assuming there are k groups, $k-1$ dummy coding variables are needed to represent each group. The evaluation is performed by leave-one-out analysis, i.e., n-fold cross-validation where n equals the number of samples. In each round, all data except one sample are used to train a model, and then we apply the model on the left-out sample to get the canonical variables. A k-Nearest Neighbor (k-NN)) classifier (Duda et al. (1973)) is applied on the canonical variable representation to predict a covariate label for the left-out sample. The accuracies of classification on original data and canonical variables will be compared to evaluate the CRSE model. The number of PLS latent variables in CRSE are chosen by nested cross-validation on the training data (i.e., not touching the test data), using the classification accuracy.

We compare the results of CRSE with that of MultiwayCCA and SCCA. Considering that the factor analysis of MultiwayCCA is too expensive and it cannot finish in a reasonable time for our data (it takes hours to finish pilot experiments on an example dataset with only 1000 variables in the expression view), we use k-means as an alternative to reduce the variable number. In analogy to the MultiwayCCA dimension reduction, we cluster all the variables of the original data into k clusters, where $k < n$, and then use the cluster centers as low-dimensional latent variables for the following procedure. Even though, MultiwayCCA still cost much more time than CRSE. Since SCCA is very slow on the full high-dimensional data, too, we use the clustered data also as input for SCCA, and for comparison purposes also with CCA. In order to prove that the clustering step has not lost much information of the original data with respect to covariate structure, we also evaluate the classification accuracy of the clustered data. Since MultiwayCCA yields by default the first shared component, we apply k-NN algorithm on the first pair of canonical variables for all the approaches.

The classification results of dataset 1 is shown in Figure 2. All the classification accuracies have a relative stable trend with the increase of k in k-NN. In Figure 2 (a) and (b), CRSE has the best performance on classification accuracy, even better than the original data. MultiwayCCA outperforms the clustered data for both views, which achieve similar classification accuracies with the original one. SCCA has the second highest accuracy in Figure 2 (a) but the worst in (b). Combining expression and metabolome with different conditions (Figure 2 (c) and (d)), CRSE achieves no better classification accuracies than the original data, but it is still the best performing method among comparisons.

MultiwayCCA, CCA and SCCA show lower accuracies than that of clustered data. The reason for the low performance of all methods compared to the original data might be the complex six-group structure of the condition covariate, which cannot be captured by a single component. This is further analyzed below.

Figure 3 shows the classification results on dataset 2, and the performance of all methods are similar with that of Figure 2. CRSE achieves the highest classification accuracy in most case, and its accuracy reaches the optimum for both expression data and metabolome data with the condition covariate information.

Since more than one canonical variables have very high correlation values, which are greater than 0.9 (see Figure 4), we use more canonical variables to do the classification-based evaluation, and the comparisons are shown in Figure 5. One, two, five and six canonical variables in CRSE are chosen to apply the k-NN algorithm, respectively. The classification accuracy is getting better and better with the increased number of latent variables. When five and six components are used, the accuracy reaches the accuracy of the original data in Figure 5(a) and it achieves a stable but higher level than that of the original data in Figure 5(b). This confirms that with a larger number of components, the relevant structure of the six-group condition covariate is successfully captured.

4 Conclusions and Discussion

Multi-view data analysis taking covariate information into account plays an important role in data mining of omics measurements and gives us a potential way to get fully aware of data structure and hidden patterns. Since there is still a lack of efficient analysis methods to address this challenge, a better understanding of the available models is needed to exploit the potentialities. In this paper we have built the pipeline of Covariate-Related Structure Extraction on paired datasets. In CRSE, we can handle high-dimensional data with small sample size and integrate covariate information into a new canonical variable space. Real datasets of *Arabidopsis thaliana* plants have been analyzed as a demonstration, and we have also shown the effectiveness of CRSE using classification-based evaluation of the extracted relationships between metabolome and gene expression variables, indicating a good separation in the canonical variable space.

CRSE achieves the highest classification accuracy in the classification-based evaluation. As for the alternative methods, SCCA has a better accuracy than CCA since it incorporates both covariate information and data structure. MultiwayCCA also has a better performance than standard CCA, but it is still time-consuming after the pre-clustering. It takes the covariate information into account but does not optimize for having significant covariate effects. The CRSE pipeline presented in this paper has the highest classification accuracy, allowing to explore the biologically relevant structure in two-view data. CRSE can be

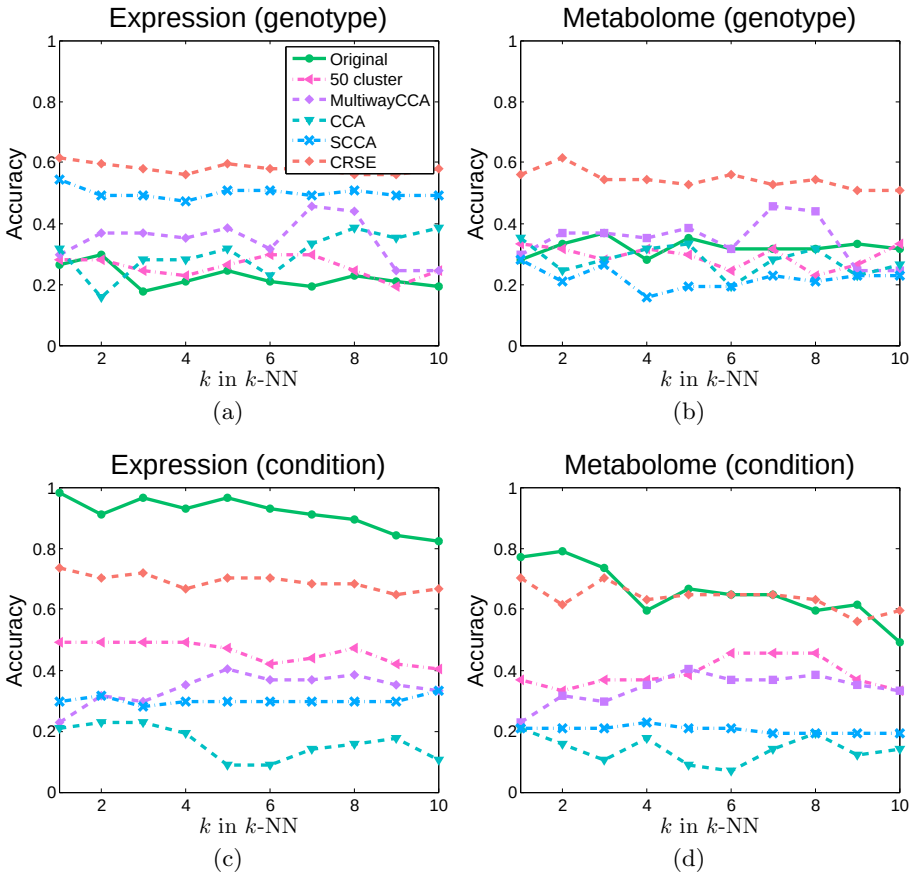

Fig. 2: Classification-based evaluation results on dataset 1 for two types of covariates: genotype and condition. Each sub-figure contains classification accuracies of original data, 50 cluster data, MultiwayCCA output, CCA output, SCCA output and canonical variable output of CRSE. (a) and (b) are the classification results for the three different genotypes, and (a) takes test samples from the expression view while (b) takes that from the metabolomic view. Classification results for the six different conditions are shown in (c) and (d), which indicate expression view and metabolomic view, respectively.

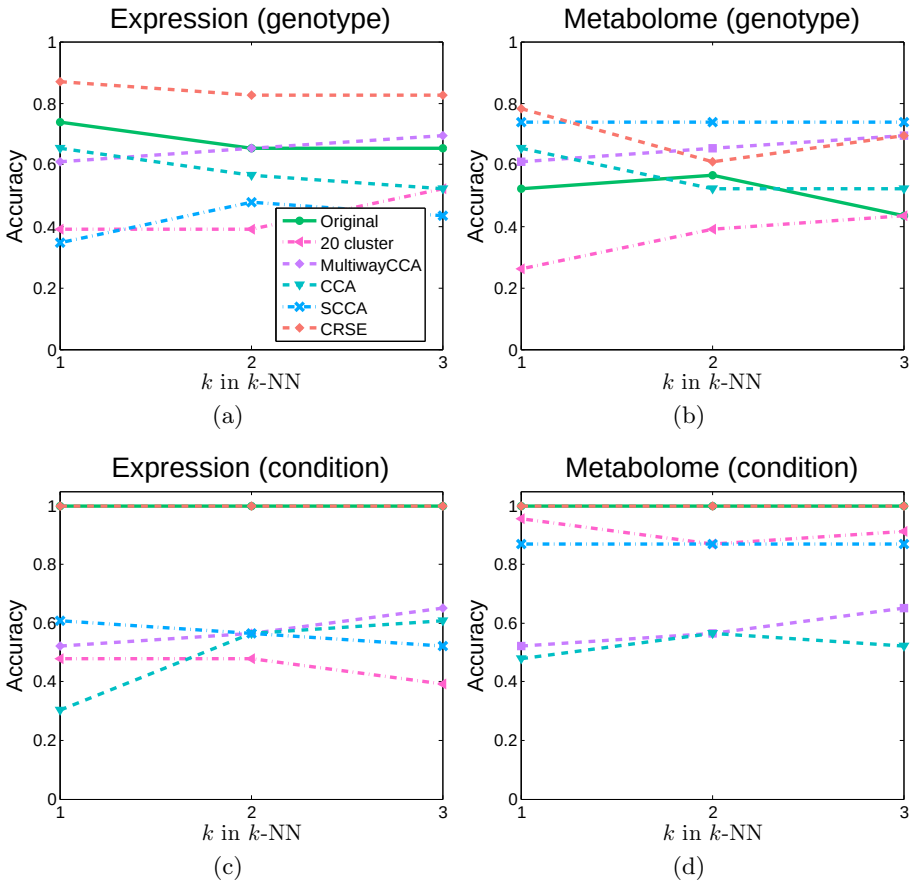

Fig. 3: Classification-based evaluation results on dataset 1 for two types of covariates: genotype and condition. Each sub-figure contains classification accuracies of original data, 20 cluster data, MultiwayCCA output, CCA output, SCCA output and canonical variables data of CRSE. (a) and (b) are the classification results for the two different genotypes, and (a) takes test samples from the expression view while (b) takes that from the metabolomic view. Classification results for the two different conditions are shown in (c) and (d), which indicate expression view and metabolomic view, respectively.

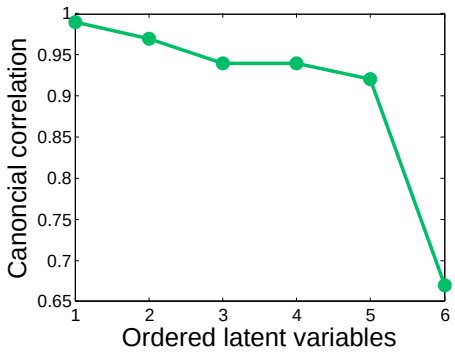

Fig. 4: The correlations of canonical variables in CRSE.

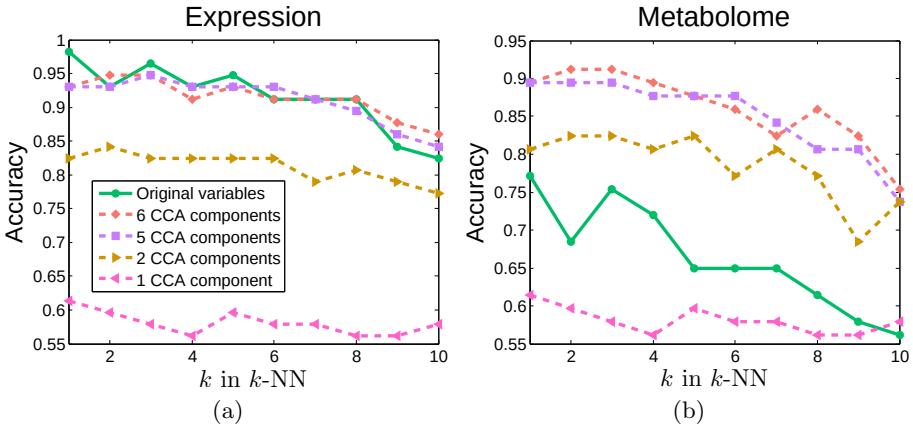

Fig. 5: Classification results using different number of latent variables. In this figure, the classification accuracies of experiment on dataset 1 under the six different conditions are shown. (a) and (b) show the classification results on the two data views, respectively.

extended to handle more than two data views by applying existing generalized CCA approaches. It should be noticed that even if two views have a high correlation in the canonical variable space obtained from the training data, especially for the first component, they do not necessarily act the same on test data, leading to differences in classification accuracy. Therefore, robustness of structure extraction approaches must be carefully examined before pursuing biological follow-up studies. In particular for complex multi-group covariates, a higher classification accuracy can be obtained if more components are chosen.

An advantage of the CRSE approach is that the dimension reduction step of the pipeline can exploit also samples available only for one data view (i.e., non-paired samples). Furthermore, it is straightforward to apply the presented approach with sparse PLS and CCA variants to improve the interpretability of components, or to replace PLS by other dimension reduction methods. From a biological perspective, it will be useful to include known gene-metabolite connections from metabolic pathways into the multi-view model and infer additional relationships for covariate structure that cannot be explained by current knowledge.

Acknowledgement

We thank Ming Jin, Jin Zhao, Basem Kanawati, Philippe Schmitt-Kopplin, Andreas Albert, J. Barbro Winkler, and Anton R. Schäffner for kindly providing the datasets used in this study.

Bibliography

Abdi, H., Williams, L. J., and Valentin, D. (2013). Multiple factor analysis: principal component analysis for multitable and multiblock data sets. *Wiley Interdisciplinary Reviews: Computational Statistics*, 5(2):149–179.

Acar, E., Gurdeniz, G., Rasmussen, M., Rago, D., Dragsted, L. O., and Bro, R. (2012). Coupled matrix factorization with sparse factors to identify potential biomarkers in metabolomics. In *IEEE 12th International Conference on Data Mining Workshops*, pages 1–8.

Acar, E., Papalexakis, E. E., Rasmussen, M. A., Lawaetz, A. J., Nilsson, M., and Bro, R. (2014). Structure-revealing data fusion. *BMC Bioinformatics*, 15(1):239.

Barkauskas, D. (2012). *FTICRMS: Programs for Analyzing Fourier Transform-Ion Cyclotron Resonance Mass Spectrometry Data*. R package version 0.8.

Boulesteix, A.-L. and Strimmer, K. (2007). Partial least squares: a versatile tool for the analysis of high-dimensional genomic data. *Briefings in Bioinformatics*, 8(1):32–44.

Choi, S. W. and Lee, I.-B. (2005). Multiblock PLS-based localized process diagnosis. *Journal of Process Control*, 15(3):295–306.

Duda, R. O., Hart, P. E., et al. (1973). *Pattern classification and scene analysis*, volume 3. Wiley New York.

Eslami, A., Qannari, E., Kohler, A., and Bougeard, S. (2014). Multivariate analysis of multiblock and multigroup data. *Chemometrics and Intelligent Laboratory Systems*, 133:63–69.

Geladi, P. and Kowalski, B. R. (1986). Partial least-squares regression: a tutorial. *Analytica Chimica Acta*, 185:1–17.

González, I., Déjean, S., Martin, P. G., Baccini, A., et al. (2008). CCA: An R package to extend canonical correlation analysis. *Journal of Statistical Software*, 23(12):1–14.

Guo, S., Ruan, Q., Wang, Z., and Liu, S. (2013). Facial expression recognition using spectral supervised canonical correlation analysis. *Journal of Computing and Information Science in Engineering*, 29(5):907–924.

Haenlein, M. and Kaplan, A. M. (2004). A beginner's guide to partial least squares analysis. *Understanding Statistics*, 3(4):283–297.

Hardoon, D. R., Szedmak, S., and Shawe-Taylor, J. (2004). Canonical correlation analysis: An overview with application to learning methods. *Neural Computation*, 16(12):2639–2664.

Horst, P. (1961). Generalized canonical correlations and their applications to experimental data. *Journal of Clinical Psychology*, 17(4):331–347.

Hotelling, H. (1936). Relations between two sets of variates. *Biometrika*, pages 321–377.

Huopaniemi, I., Suvitaival, T., Nikkilä, J., Orešič, M., and Kaski, S. (2010). Multivariate multi-way analysis of multi-source data. *Bioinformatics*, 26(12):i391–i398.

Jamali, M. and Ester, M. (2010). A matrix factorization technique with trust propagation for recommendation in social networks. In *Proceedings of the 4th ACM Conference on Recommender Systems*, pages 135–142. ACM.

Jiang, M., Cui, P., Liu, R., Yang, Q., Wang, F., Zhu, W., and Yang, S. (2012). Social contextual recommendation. In *Proceedings of the 21st ACM International Conference on Information and Knowledge Management*, pages 45–54. ACM.

Klami, A., Virtanen, S., and Kaski, S. (2013). Bayesian canonical correlation analysis. *Journal of Machine Learning Research*, 14(1):965–1003.

Krzanowski, W. (1984). Principal component analysis in the presence of group structure. *Applied Statistics*, pages 164–168.

Lanckriet, G. R., De Bie, T., Cristianini, N., Jordan, M. I., and Noble, W. S. (2004). A statistical framework for genomic data fusion. *Bioinformatics*, 20(16):2626–2635.

Lee, C. M., Mudaliar, M. A., Haggart, D., Wolf, C. R., Miele, G., Vass, J. K., Higham, D. J., and Crowther, D. (2012). Simultaneous non-negative matrix factorization for multiple large scale gene expression datasets in toxicology. *PLoS ONE*, 7(12).

Luo, Y., Tao, D., Ramamohanarao, K., Xu, C., and Wen, Y. (2015). Tensor canonical correlation analysis for multi-view dimension reduction. *IEEE Transactions on Knowledge and Data Engineering*, 27(11):3111–3124.

Pinheiro, J. C. and Bates, D. M. (2000). Linear mixed-effects models: basic concepts and examples. *Mixed-effects Models in S and S-Plus*, pages 3–56.

Ritchie, M. E., Phipson, B., Wu, D., Hu, Y., Law, C. W., Shi, W., and Smyth, G. K. (2015). limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Research*, 43(7):e47.

Saunders, C., Gammerman, A., and Vovk, V. (1998). Ridge regression learning algorithm in dual variables. In *Proceedings of the 15th International Conference on Machine Learning*, pages 515–521. Morgan Kaufmann.

Smilde, A. K., Westerhuis, J. A., and de Jong, S. (2003). A framework for sequential multiblock component methods. *Journal of Chemometrics*, 17(6):323–337.

Smyth, G. K. (2004). Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Statistical Applications in Genetics and Molecular Biology*, 3(1):1–25.

Svante, W., Sven, H., Torbjrn, L., and Michael, S. (1987). PLS modeling with latent variables in two or more dimensions. *Partial Least Squares Model Building: Theory and Application*.

Sweeney, K. T., McLoone, S. F., and Ward, T. E. (2013). The use of ensemble empirical mode decomposition with canonical correlation analysis as a novel artifact removal technique. *IEEE Transactions on Biomedical Engineering*, 60(1):97–105.

Tenenhaus, M. and Vinzi, V. E. (2005). PLS regression, PLS path modeling and generalized procrustean analysis: a combined approach for multiblock analysis. *Journal of Chemometrics*, 19(3):145–153.

Vía, J., Santamaría, I., and Pérez, J. (2007). A learning algorithm for adaptive canonical correlation analysis of several data sets. *Neural Networks*, 20(1):139–152.

Vinod, H. D. (1976). Canonical ridge and econometrics of joint production. *Journal of Econometrics*, 4(2):147–166.

Wendorf, C. A. (2004). Primer on multiple regression coding: Common forms and the additional case of repeated contrasts. *Understanding Statistics*, 3(1):47–57.

Westerhuis, J. A., Kourti, T., and MacGregor, J. F. (1998). Analysis of multiblock and hierarchical PCA and PLS models. *Journal of Chemometrics*, 12(5):301–321.

Witten, D. M., Tibshirani, R., and Hastie, T. (2009). A penalized matrix decomposition, with applications to sparse principal components and canonical correlation analysis. *Biostatistics*, 10(3).

Witten, D. M. and Tibshirani, R. J. (2009). Extensions of sparse canonical correlation analysis with applications to genomic data. *Statistical Applications in Genetics and Molecular Biology*, 8(1):1–27.

Wold, S., Kettaneh, N., and Tjessem, K. (1996). Hierarchical multiblock PLS and PC models for easier model interpretation and as an alternative to variable selection. *Journal of Chemometrics*, 10(5-6):463–482.

Zhou, G., Cichocki, A., Zhang, Y., and Mandic, D. (2015). Group component analysis for multiblock data: Common and individual feature extraction. *IEEE Transactions on Neural Networks and Learning Systems*, PP(99):1–14.